적응 프루닝 알고리즘과 PDT-SSS 알고리즘을 이용한 한국어 연속음성인식에 관한 연구

황철준[†] · 오세진[††] · 김범국[†††] · 정호열[††††] · 정현열[†††††]

요 약

연속음성인식 시스템의 실용화를 위해서 가장 중요한 것은 높은 인식 성능을 가지면서 동시에 실시간으로 인식되어야 한다. 이를 위하여 본 연구에서는 먼저 연속음성인식의 인식률 향상을 위하여 효과적인 음향모델을 구성하기 위하여 PDT-SSS(Phonetic Decision Tree-based Successive State Splitting) 알고리즘을 도입하여 HM-Net을 구성하고, 언어모델로서 반복학습을 이용하여 인식률 향상을 제고한다. 그리고, 기존의 연구에서 유효함이 입증된 프레임 단위 적응 프루닝 알고리즘을 연속음성에 적용하여 인식 속도를 개선하고자 한다. 제안된 방법의 유효성을 확인하기 위하여, 남성 4인이 항공편 예약 관련 음성에 대하여 인식 실험을 수행하였다. 그 결과 연속음성인식률 90.9%, 단어인식률 90.7%의 높은 인식성능을 얻었으며, 적응 프루닝 알고리즘을 적용한 경우 인식성능의 저하없이 약 1.2초(전체의 15%)의 인식시간을 줄일 수 있어 제안된 방법의 유효성을 확인할 수 있었다.

A Study on the Korean Continuous Speech Recognition using Adaptive Pruning Algorithm and PDT-SSS Algorithm

Chul-Joon Hwang[†], Se-Jin Oh[††], Bum-Koog Kim[†††], Ho-Youl Jung[††††] and Hyun-Yeol Chung[†††††]

ABSTRACT

Efficient continuous speech recognition system for practical applications requires that the processing be carried out in real time and high recognition accuracy. In this paper, we study the acoustic models by adopting the PDT-SSS algorithm and the language models by iterative learning so as to improve the speech recognition accuracy. And the adaptive pruning algorithm is applied to the continuous speech. To verify the effectiveness of proposed method, we carried out the continuous speech recognition for the Korean air flight reservation task. Experimental results show that the adopted algorithm has the average 90.9% for continuous speech recognition and the average 90.7% for word recognition accuracy including continuous speech. And in case of adopting the adaptive pruning algorithm to continuous speech, it reduces the recognition time of about 1.2 seconds(15%) without any loss of accuracy. From the result, we proved the effectiveness of the PDT-SSS algorithm and the adaptive pruning algorithm.

1. 서 론

음성은 인간이 사용하는 가장 기본적인 의사소통을 위한 수단이며, 편리함과 경제성의 측면에서 다른 방법에 비해 우수한 특성을 가진다. 최근 컴퓨터 하드웨어 기술의 급속한 진보와 음성처리 기술의 발전으로 인하여 음성인식의 실용화가 실질적인 문제로서 관심이 증대되고 있다. 이러한 관심이 증대되면서, 음성인식에 관한 연구는 실용화에 초점이 모아지면서 최근 몇 년간의 눈부시게 발전하여 일부 태스크에서 상업용 시스템이 구현되고 있는 실정이다[1,2].

이 논문은 1998년도 한국학술진흥재단 대학부설연구소과제(과제번호 98-005-E00017) 연구비에 의해 연구되었음

† 정회원, 대구과학대학 정보전자통신계열 전임강사
†† 대구과학대학 정보전자통신계열 전임강사
††† 대구과학대학 정보전자통신계열 조교수
†††† 정회원, 영남대학교 전자정보공학부 조교수
††††† 영남대학교 전자정보공학부 교수

일반적으로 음성인식 시스템의 실용화를 위해서 가장 중요한 것은 높은 인식 성능을 가지면서 동시에 실시간으로 인식되어야 할 필요가 있으나, 이 두 요구조건은 상충되는 사항이다. 예를 들어, 인식시간을 줄이기 위해 탐색 공간을 대량으로 프루닝하면서 간단한 음향학적 모델을 사용하면 인식 속도를 쉽게 향상시킬 수는 있지만 이에 따르는 인식률의 저하는 피할 수 없다. 고립 단어 단위에서는 인식률을 향상시키거나 혹은 그대로 유지하면서 인식 속도를 높이는 것에 대해서는 어느 정도 연구 성과가 있지만, 대용량 어휘를 대상으로 하는 연결음성인식 또는 연속음성인식에서는 아직까지 많은 연구가 필요하다. 실제로 이용할 수 있는 실용화 시스템을 구축하기 위해서는 높은 인식성능과 빠른 인식속도의 두 조건을 동시에 만족하지 않으면 안 된다[3].

고립단어 인식에 있어서는 약간의 잡음이 있는 환경하에서도 95%이상의 인식 성능을 가지며, 한정된 태스크 범주내의 연속음성인식에서도 90%이상의 높은 인식률을 가진 시스템이 많이 개발되고 있으며, 인식 태스크를 확장하기 위한 여러 가지 연구들이 진행되고 있다[4-6]. 국외에서는 Name Dialing System, 증권 거래 시스템 등과 같이 고립 단어를 대상으로 하는 수종의 시스템이 개발되어 실용화되고 있으며, Dragon Dictate, 날씨 안내 시스템 등과 같이 한정된 태스크에서의 연속음성 인식에서도 거의 실시간으로 동작하는 시스템이 많이 개발되어 실용화 단계에 있다[7,8]. 또한 자연발화(Natural Speech) 인식에 대한 연구도 활발하게 진행 중에 있다. 국내의 경우에 있어서는 최근의 음성인식에 대한 관심의 증대로 인하여 증권 안내 시스템, 부서 안내 시스템 등과 같이 고립단어를 대상으로 하는 인식 시스템이 개발되어 실제 상용화되고 있지만, 대어휘를 대상으로 하는 실시간 음성인식 시스템 구현을 위한 고속화에 대한 연구는 아직까지 많이 부족한 실정이다[1,2].

기존의 연구에서 음성인식 시스템의 실용화를 위해 높은 인식 성능뿐만 아니라 빠른 인식 속도를 가지는 시스템을 구축하기 위하여 연구를 진행해 왔다[9-13]. 그 연구 결과로서 개발된 음성인식 기능을 가진 주소 인식시스템으로, 한국어 주소의 특징을 고려하여 연결단어 인식을 태스크로 하고 있다. 인식시간의 경우, 기존에 제안한 프레임 단위 적응 프루닝 문턱치 알고리즘을 적용하여 탐색 공간이 효과적으로 줄어듦을 확인하였다.[12,13]. 그러나, 주소음성이 비록 대용량이긴 하지만, 하나의 상위 행정단위를 인식하고 인식된 행정단위의 하위 행정단위만을 인식 대상으로 하기 때문에 고립단어인식과 비슷한 과정을 거치게 된다.

따라서 본 논문에서는 프레임 단위 적응 프루닝 문턱치 알고리즘을 연속음성인식에 적용하여 그 유효성을 확인하고자 한다. 먼저 연속음성의 인식률을 향상시키기 위하여 HM-Net을 도입하였다. HM-Net은 비슷한 파라미터를 가지는 HMM의 상태와 출력 확률분포를 하나로 하여 상향(bottom-up)으로 공유하는 방법이고, 이를 생성하기 위하여 모든 음소모델에 대응하는 상태공유를 자동으로 결정하는 SSS 알고리즘을 이용하여 작은 상태에 보다 정확한 문맥의존 모델을 생성하였다. 언어모델로는 반복학습을 통한 N-gram을 도입하여 인식률 향상을 도모하였다[14]. 이렇게 얻은 시스템에 프레임 단위 적응 프루닝 문턱치 알고리즘을 적용하여 연속음성인식에서의 높은 인식률과 빠른 인식속도를 가지는 시스템 개발을 연구의 대상으로 하고자 한다.

논문의 구성은 다음과 같다. 2장에서는 한국어 음성학적 규칙에 대하여 설명하고, 3장에서는 HM-Net을 구성하기 위한 SSS 알고리즘과 음소결정트리를 도입한 PDT-SSS 알고리즘에 대하여 설명하고, 4장에서는 프레임 단위 적응 프루닝 문턱치에 대하여 소개한다. 5장에서는 전체 시스템의 구성과 인식실험 방법, 6장에서는 인식실험 결과를 기술한 다음, 마지막으로 7장에서 본 논문의 결론을 맺는다.

2. 한국어 음성학적 규칙

한국어에는 다른 언어와는 달리 많은 문법과 음운 규칙이 있다. 본 연구에서는 한국어에 적합한 문맥의존 음향모델을 작성하기 위해 결정트리 기반 SSS 알고리즘의 상태분할에서 음소 질의어 집합의 구성에 한국어 음성학적 지식[15]을 이용하였다. 본 연구에서 적용한 음성학적 규칙을 표 1에 나타내었다.

표 1에 나타낸 것과 같이 적용한 규칙은 크게 모음, 자음, 유성음, 비음, 유음, 반모음과 묵음으로 나눈다. 이 중에서 모음은 혀의 위치, 입의 크기, 혀의 높이, 좁힘점위치, 좁힘점간극 등과 같이 크게 5부분으로 분류하였다. 그리고 자음은 조음자리, 조음방법

표 1. 한국어 음성학적 규칙

유성음				자음	조음자리		양순음
모 음							치(조)음
모음	혀위치	전설	비원순				경구개음
			원순				연구개음
		중설	비원순				성문음
			원순		조음방법	파열음	무기연음
		후설	비원순				유기경음
			원순				무기경음
	입크기	협(狹)				파찰음	무기연음
		반협(半狹)					유기경음
		광(廣)					무기경음
	혀높이	고모음				마찰음	무기연음
		중고모음					무기경음
		중저고음			비 음		
		저모음			유 음		
	좁힘점 위치	경구개음			반모음		
		연구개음					
		인두음					
	좁힘점 간극	폐모음					
		반폐모음					
		개모음		묵 음			

등과 같이 2부분으로 분류하고, 조음방법의 경우 파열음, 파찰음, 마찰음으로 다시 나누었다. 본 연구에는 음성학적 규칙을 문맥의 좌, 우를 포함하여 총 162 부분으로 분류하였으며, 이를 이용하여 음소 질의어 집합을 작성하였다. 이렇게 작성한 음소 질의어 집합을 결정트리에 의한 상태분할에 사용하였다.

3. HM-Net과 PDT-SSS 알고리즘

3.1 Hidden Markov Network(HM-Net)

SSS 알고리즘에 의해 작성한 HM-Net은 여러 개의 상태를 연결한 네트워크로 표현되며, HM-Net의 각 상태는 상태번호, 허용할 수 있는 문맥 클래스, 선행음소와 후행음소 리스트, 자기천이확률과 후행상태로의 천이확률 그리고 출력확률분포 파라미터와 같은 정보를 포함한다. HM-Net에서는 문맥정보가 주어지면, 이 문맥을 허용할 수 있는 상태를 선행상태와 후행상태 리스트의 제약 내에서 연결하여 이 문맥에 대한 모델을 결정할 수 있다. 이 모델은 자기천이와 다음 상태로의 천이만을 고려한 left-to-right 모델로 간주할 수 있으므로 일반적인 HMM과 같이 Baum-Welch 알고리즘에 의해 파라미터를 추정할 수 있다.

3.2 SSS 알고리즘

SSS(Successive State Splitting) 알고리즘[16,17]은 모든 문맥을 나타내는 1상태의 초기모델로부터 문맥방향과 시간방향으로 상태분할 후 자동적으로 HM-Net[16,17]의 구조를 결정하는 알고리즘이다. SSS 알고리즘으로 HM-Net을 작성하는 단계를 그림 1에 나타내었다.

전체적으로 간략히 설명하면 다음과 같다. 우선 유사음소단위(PLUs)를 기본단위로 모든 모델을 연결한 네트워크 구조의 초기모델로서 각각의 모델은 하나의 상태와 그 상태를 시단에서 종단까지 결합하여 전체 학습 데이터로부터 작성한다. 상태의 분할은 경로분할을 동반하는 문맥방향과 경로분할을 동반하지 않는 시간방향이 있는데, 출력확률의 likelihood에 따라 한 방향으로만 수행된다. 문맥방향으로 분할할 때는 경로분할에 동반된 각각의 경로에 할당된 문맥 클래스도 동시에 분할된다. 따라서 문맥 클래스의 분할에 포함된 모든 상태 중에서 학습 데이터에 대한 누적 likelihood 확률이 가장 큰 쪽의 상태를 분할하도록 선택된다. 시간방향으로의 상태분할에서도 누적 likelihood 확률이 높은 쪽 상태를 분할하도록 선택된다. 이상의 상태분할을 반복하여 HM-Net의 구조가 결정된다.

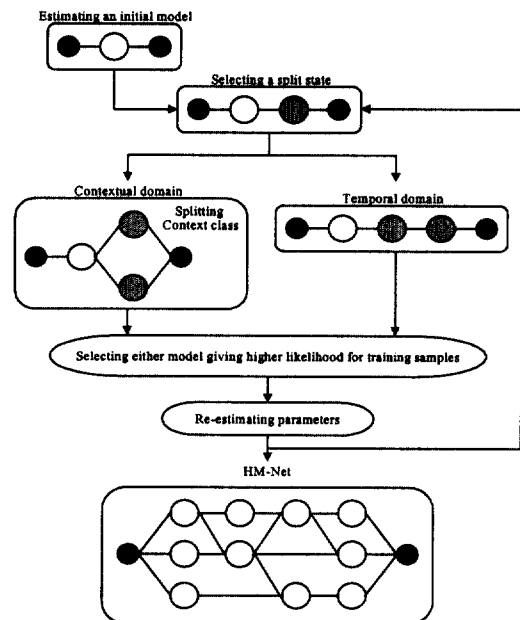



그림 1. SSS 알고리즘의 구성도

3.3 PDT-SSS 알고리즘

음소결정트리[18]는 음소의 음향적 변동을 파악하는 것으로, 미지 음소환경의 음향적 특성을 예측하는 방법이다. 음소결정트리는 뿌리(root)를 음소환경에 독립한 2진 트리로 나타내고 뿌리에서 잎 방향으로 문맥클래스의 분할을 수행한다(그림 2). 이 트리는 뿌리에서 잎 방향으로 진행함에 따라 음소환경의 의존도가 강한 단위를 나타내는 계층적 구조를 가지며, 일반적으로 잎 부분에 모델을 대응시키게 된다. 트리의 각 노드에서는 경험적으로 음소유사성에 기인한 질의어를 할당하여 yes와 no에 의해 문맥클래스를 두 개로 분할한다. 음소환경과 음소군에 따라서 각 질의어를 구성한다. 이러한 음소환경을 트리의 뿌리 노드에서 질의어를 찾아 반드시 잎에 대응시키기 위해, 미지의 음소환경에서 음향학적으로 가장 유사한 잎의 노드로 분류된다고 할 수 있다. 이를 위해, 출현하지 않는 음소환경을 음소환경 독립모델 등으로 대치할 필요도 있다.

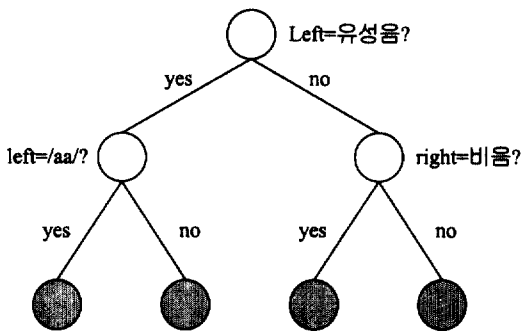



그림 2. 음소결정트리

본 연구에서는 SSS 알고리즘에 한국어 음성학적 지식으로 구성한 음소질의어 집합을 이용한 음소 결정트리에 기반한 상태분할 방법을 도입한 PDT-SSS (Phonetic Decision Tree-based SSS) 알고리즘[19]을 이용하였다. PDT-SSS는 SSS 알고리즘의 문맥방향 상태분할에 음소 결정트리를 결합한 것으로 HM-Net에서 새로운 상태의 모델 파라미터 공유와 학습데이터에 출현하지 않는 미지의 문맥에 대한 학습을 수행할 수 있도록 구성되어 있다. 여기서 음소결정트리는 2진 트리로서 각 노드는 음소질의어로 구성되어 있다. 각 음소모델의 공유 파라미터는 각 트리의 잎(leaf) 노드와 연관되고, 문맥의존 모델은 음소 질의어에 의해 트리의 뿌리(root) 노드에서 잎 노드까지 조사하여 임의의 문맥에 할당된다. PDT-SSS의 특징은 허용할 수 있는 문맥 클래스는 음소 질의어에 따른 결정트리에 의해 분할된다는 것이다. 또한, 하나의 상태가 분할될 때, 두 개의 혼합수는 새로운 상태와 관련된 것이 아니고 새로운 상태에 대한 단일 가우스 분포는 학습 샘플로부터 계산된다. 따라서, PDT-SSS 알고리즘이 적절한 문맥 클래스의 분할과 임의의 문맥을 표현할 수 있기 때문에 보다 정확한 HM-Net을 작성할 수 있게 된다. PDT-SSS 알고리즘의 주요 내용은 다음과 같다.

1) 한국어 음성학적 지식에 의한 음소 질의어 집합을 작성한다.
2) Baum-Welch 알고리즘으로 초기 HM-Net을 학습한다.(각 상태는 단일 가우스 분포)
3) SSS 알고리즘과 같이 식 (1)에 의해 최적 분포를 가지는 상태를 선택한다.
4) 문맥방향과 시간방향으로 분할할 상태를 선택한다.
   • 각 음소 질의어에 대해 문맥방향으로 분할할 때,
     i) 질의어에 대해 허용할 수 있는 문맥 클래스의 분할과 두 개의 단일 가우스 분포를 추정한다.(각 가우스 분포는 yes 또는 no에 해당)
     ii) 새로운 상태에 각 문맥 클래스와 각 가우스 분포를 할당한다.
   • 각 음소 질의어에 대해 시간방향으로 분할할 때,
     i) Baum-Welch 재추정에 의해 두 개의 단일 가우스 분포를 추정한다.
     ii) 새로운 상태에 각 가우스 분포를 할당하고 문맥 클래스를 복사한다.
5) 학습 샘플의 likelihood에 근거하여 문맥방향과 시간방향에서 최적의 HM-Net을 선택한다.
6) Baum-Welch 알고리즘에 의해 HM-Nets의 상태를 재학습한다.
7) 미리 정의한 상태수에 도달할 때까지 단계 3부터 반복한다.

단계 3에서 분할될 상태의 선택은 식(1)에 의해 계산된다.

$$d_i = n_i \sum_{p=1}^{P} \frac{\sigma_{ip}^2}{\sigma_{Tp}^2} \tag{1}$$

여기서, $\sigma_{ip}^2, \sigma_{Tp}^2$는 상태 i의 분포 분산과 모든 샘플의 분산(정규화 계수)을 나타내고, n_i는 상태 i의 추정에 이용한 음소 샘플의 수를, P는 특징 벡터의 차원 수를 각각 나타낸다.

4. 프레임 단위 적응 프루닝 알고리즘

4.1 고정 프루닝 알고리즘

음성인식을 수행하기 위해서는 출력확률 계산과 탐색의 2가지 계산과정을 필요로 한다. HMM을 이용한 음성인식에서의 출력확률 계산은 임의의 한 시점에서 관측된 음성을 출력하는 주어진 HMM의 상태의 확률계산이며, 탐색은 주어진 음성 입력에 대한 최상의 상태열을 구하는 문제로 볼 수 있다. 이러한 탐색에 소요되는 시간은 음향학적 모델의 복잡성에 의해서는 크게 영향을 받지 않으나, 인식대상의 규모에 따른 영향은 크다. 즉, 인식에 있어서 모든 가능한 상태열들을 고려할 경우, 입력된 음성에 대한 최고 likelihood의 상태 열(단어, 문장)을 찾기 위한 탐색공간은 지수 함수적으로 증가한다.

현재까지 대부분의 시스템에서는 프레임 동기형의 빔 탐색법을 이용하고 있는데 이 방법은 각 후보의 likelihood를 비교하고 상위 일정 개수(문턱치 이하의 것)에 대해서만 후속 정합을 고려하는 방법으로 다음과 같이 나타낸다[20].

$$D_{\min}(i,j) \le D_{\min}(i,j^*) + \delta \tag{2}$$

이 방법은 i프레임에서의 최적의 경로 (i,j^*)에 대해 문턱치 δ이내의 상위 몇 개(빔 폭)만을 후속탐색에서 고려하고 나머지는 탐색으로부터 제외하는 방법이다. 정합은 입력 프레임과 식 (2)의 범위내의 노드에 대응하는 음향 모델과의 정합을 의미한다. 여기서 각 노드의 likelihood를 비교하여 상위 일정 개수를 선택한 후, 여기서부터 전개되어지는 노드들과 입력 $i+1$프레임과 정합한다.

탐색공간을 더욱 제한하는 방법으로써 프루닝 기법[21,22]이 있다. 이 방법은 각 프레임에 있어서 최대 likelihood를 $g_{\max}$로 하고, $g_{\max} = -\lambda$ (λ는 여유분을 둔 문턱치) 에 만족하지 않는 후보에 대해서는 그 시점 이후의 탐색을 프루닝함으로써 탐색공간을 감소시킨다.

먼저 One-pass Viterbi 알고리즘의 누적대수 likelihood 확률 $p_q^n(i,j)$의 i프레임에 대한 최대치 $P_{\max}(i)$을 다음과 같이 구할 수 있다[23].

$$P_{\max}(i) = \max_{j,n,q} P_q^n(i,j) \tag{3}$$

이렇게 구해진 최대 likelihood에 대해 식 (4)과 같은 조건을 만족하는 각 상태 q의 각 단어(또는 PLU)에 대해서만 탐색을 수행하고 나머지는 제외하는 기법을 다음 식으로 나타낼 수 있다.

$$\max_j P_q^n(i,j) < P_{\max}(i) - \lambda \tag{4}$$

빔 탐색법에서 가장 중요한 것은 각 후보의 likelihood의 정도이다. 정도가 낮은 경우, 정해로 얻어진 후보가 프루닝에 의해 제외되는 오류가 있을 수 있다. 즉, 어떤 시점(처리 프레임)에서 그 노드까지의 누적 likelihood가 크지 않을 경우 정해가 될 수 있음에도 불구하고 탐색에서 제외되어 최적성을 보장받지 못하게 되므로, 빔 폭의 제한과 프루닝조건을 엄격하게 함으로써 최적해를 잃을 우려가 있다. 따라서, 인식정도에 영향을 주지 않기 위해서는 빔 폭과 프루닝조건을 완화시키면서 탐색공간을 감소시키는 방법을 찾을 필요가 있다.

4.2 프레임 단위 적응 프루닝 알고리즘

앞 절에서 설명한 방법이 각 프레임에서 후보 단어들을 이전에 제안된 방법들에 비해 보다 효과적으로 제한할 수 있었지만, 여전히 탐색할 필요가 없는 공간을 탐색한다. 따라서 여기서는 인식 과정 중에 탐색 공간을 효과적이고 자동으로 줄이기 위하여 프레임 단위 적응 프루닝 알고리즘을 제안한다.

이 알고리즘은 이웃 프레임사이의 최대 likelihood 확률들의 상관성이 크므로 앞 프레임의 최대 likelihood 확률로부터 효과적인 프루닝 문턱치를 얻을 수 있다는 점에 착안하여, 앞 프레임의 최대 likelihood 확률과 후보 likelihood 확률들의 조합으로 현재 프레임에서의 프루닝 문턱치를 프레임 단위로 갱신하는 방법이다.

현재 프레임의 프루닝 문턱치는 식 (5)을 이용하여 계산된다.

$$\lambda(k)=\frac{1}{N}\sum_{s=1}^{N}\{P_{\max}(i-1,j*)-P_{hyp}(i-1,s)\} \quad (5)$$

여기서, $P_{\max}(i-1,j*)$는 프레임 $i-1$에서 최대 likeli -hood 확률이고, $P_{hyp}(i-1,s)$는 프레임 $i-1$에서 여러 후보들의 likelihood 확률이고, 그리고 N은 프레임 $i-1$에서 후보의 수이다.

식 (5)로부터 알 수 있는 바와 같이 제안된 알고리즘은 현재의 문턱치가 인식 과정 중에 얻어질 수 있기 때문에, 인식 태스크가 바뀌더라도 문턱치를 구하기 위하여 여러 번의 사전 실험을 필요로 하지 않는다. 또한, 문턱치가 적응적으로 얻어지기 때문에 다른 환경 하에서도 인식 속도를 향상시킬 수 있다.

5. 인식 실험

한국어 음성학적 지식과 결정트리 기반 상태분할 알고리즘에 의해 작성한 한국어 문맥의존 음향모델의 유효성을 확인하기 위해 음소, 단어 및 연속음성 인식 실험을 수행하였다. 그리고, 기존에 연결단어 음성에 적용했던 적응 프루닝 문턱치 알고리즘의 유효성을 확인하기 위해 연속음성을 대상으로 인식실험을 수행하였다. 그림 3에 인식시스템의 전체 구성도를 나타내었다.

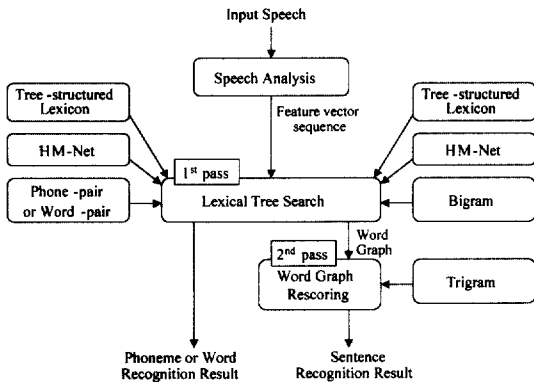



그림 3. 음성인식 시스템의 전체 구성도

음소 및 단어 인식실험에서는 문맥의존 음향모델을 작성하기 위해 사용된 음성데이터는 452단어를 38명이 2회 발성한 국어공학센터(KLE)의 음성 데이터베이스를 사용하였다. 이를 본 연구에서는 2 부분으로 나누어 학습과 평가에 사용하였다. 첫 번째 35명이 1회 발성한 15,820단어를 문맥의존 음향모델을 학습하는데 사용하였으며, 두 번째 학습에 참가하지 않은 3명이 첫 번째 발성한 1,356단어를 화자독립 평가에 각각 사용하였다. 연속음성 인식실험의 경우에는, 국어공학센터(KLE)의 단어음성과 본 연구실의 항공편 예약관련 200문장(YNU200) 연속음성 데이터베이스를 사용하였다. 음향모델의 학습을 위해 452 단어를 35명이 1회 발성한 15,820단어와 200문장을 8명이 1회 발성한 1,600문장을 문맥의존 음향모델을 학습하는데 사용하였으며, 학습에 참가하지 않은 4명의 200문장을 화자독립 연속음성인식 평가에 사용하였다.

모든 음성데이터는 16kHz의 샘플링과 16bits로 양자화 되었으며, $1-0.97z^{-1}$의 전달함수로 프리엠퍼시스 하였으며, 25ms의 해밍 윈도우를 곱하여 10ms씩 이동하면서 분석하였다. 이를 통해 음성 특징 파라미터는 12차 LPC-멜 켑스트럼 계수와 정규화된 대수 에너지에 1차 및 2차의 차분 성분을 포함하여 총 39차의 특징 파라미터를 구하였다. 표 2에 음성데이터 및 분석조건을 나타낸다.

또한, PDT-SSS 알고리즘에 의한 문맥방향의 상태분할을 위해 162개(문맥의 좌, 우)의 음소 질의어 집합을 한국어 음성학적 지식에 근거하여 작성하였다. 초기 HM-Net의 구조는 48개의 유사음소단위를 병렬로 연결하여 141개의 상태를 가지도록 구성하였다. 모든 HM-Net은 혼합수 4를 가지며 200에서 1,200상태까지는 200상태씩 증가시켰으며, 상태수 2,000인 HM-Net도 학습하였다.

표 2. 음성 데이터 및 분석조건

음성 데이터				
발성형태	단어	단어	문장	문장
화자	남성 35	남성 3	남성 8	남성 4
사용단계	학습	인식	학습	인식
단어(문장)수	452	452	200	200
발성횟수	1	1	1	1
발성환경	방음부스			
분석조건				
샘플링 주파수	16khz			
분해능	16bits			
Pre-emphasis	$1-0.97z^{-1}$			
Window	Hamming window(25msec)			
분석 주기	10msec			
특징 파라미터 (39 차)	MFCC(12) + Power(1) + ΔMFCC(12) + ΔPower(1) + ΔΔMFCC(12) + ΔΔPower(1)			

음소 및 단어인식 알고리즘은 One-Pass Viterbi beam 탐색 알고리즘[24,25]으로서 음소인식의 경우 한국어 음소제약을 가지는 phone-pair 문법을, 단어인식의 경우 word-pair 문법을 각 사용하였다.

연속음성 인식 알고리즘은 Multi-pass 탐색 알고리즘[24]으로서 1-pass 탐색의 경우, 단어 2-gram 언어모델을 이용하여 프레임 동기형 Viterbi beam 탐색을 수행한 후 단어 그래프를 출력한다. 2-pass 탐색의 경우 1-pass의 단어 그래프와 보다 정밀한 단어 3-gram을 이용하여 A* stack decoding 탐색을 수행한 후 인식결과를 출력한다.

6. 인식 실험 결과

제안된 알고리즘의 유효성을 확인하기 위하여 수행된 인식실험 결과를 나타낸다. 먼저 그림 4에 화자독립 음소인식 실험결과를, 그림 5에 화자독립 단어인식 실험결과를 각각 나타내었다.

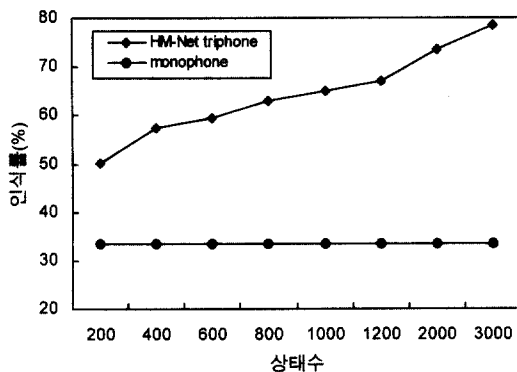



그림 4. 화자독립 음소인식 실험결과

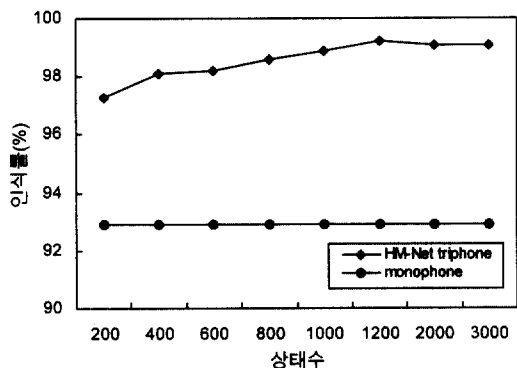



그림 5. 화자독립 단어인식 실험결과

그림 4의 음소인식률의 경우, 모노폰에 대해 KLE 3인 평균 33.5%를 나타내고 있다. 그리고 HM-Net triphone에 대해서는 상태수 200일 때 평균 50.2%, 상태수 3,000일 때 평균 78.6%를 나타내고 있다. 모노폰과 상태수 3,000일 때의 HM-Net triphone을 비교하면 HM-Net을 이용한 경우가 평균 45.1%의 음소인식률 향상을 보였다. 또한 상태수 200과 상태수 3,000일 때의 HM-Net을 비교하면 평균 28.4%의 인식률 향상을 보이고 있다. 마찬가지로 그림 5의 단어인식률에서도 모노폰의 경우 KLE 3인 평균 92.9%, 상태수 200일 때 평균 97.3%, 상태수 3,000일 때 평균 99.1%의 평균 단어인식률을 구하였다. 그림 5에서도 모노폰과 HM-Net을 비교하면 상태수 3,000일 때의 HM-Net을 이용한 경우가 평균 6.2%의 인식률 향상을 나타내고 있다.

그리고 본 연구에서는 한국어의 다양한 특성을 고려하여 48개의 유사음소단위(PLUs)를 사용하였는데, 만약 48개의 유사음소단위로 triphone을 작성한다면, 실제 음성인식 시스템에서 110,592(48^3)개의 triphone을 만들어야 하지만 실제로 많은 수의 triphone을 작성하여 인식 시스템에 사용하면 계산적 부하가 발생한다. 실제 본 연구에서 학습에 사용된 음성 데이터에 출현하는 음소단위로 생성될 수 있는 triphone의 수는 2,164개이지만 PDT-SSS 알고리즘에 의한 문맥방향으로 한국어 음성학적 기식과 결정트리기반 상태분할을 수행한 결과 108,289개의 HM-Net triphone을 작성할 수 있었고, 유사한 확률을 가지는 상태를 공유하여 시스템의 계산적 부하를 최소화하였다.

그림 6에 상태수의 변화에 따른 화자독립 연속음성인식률을 나타내고, 그림 7에 인식 문장에 포함된 단어인식률을 각각 나타내었다.

그림 6에서 상태수 1,000일 때 HM-Net triphone의 경우 1-pass의 인식률은 평균 86.9%로서 단일 HMM에 비해 평균 9.9%의 인식률을 향상을 보이고, 상태수 800일 때 HM-Net triphone의 경우 2-pass의 인식률은 평균 90.9%로서 단일 HMM에 비해 평균 4.1%의 인식률을 향상을 보였다. 또한 그림 7에서 인식대상인 연속음성에 포함된 798단어에 대한 인식률은 상태수 1,000일 때 HM-Net triphone의 경우 1-pass 인식률은 평균 89.9%로서 단일 HMM에 비해 평균 7.6%의 인식률 향상을 보이고, 상태수 200일

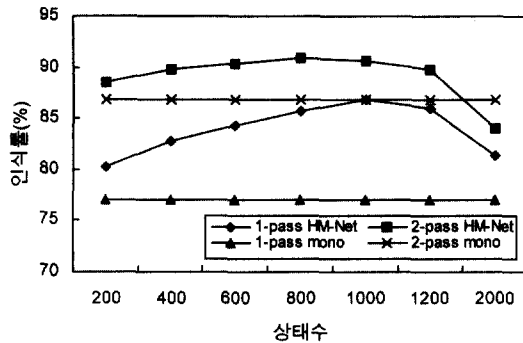



그림 6. 화자독립 연속음성인식률

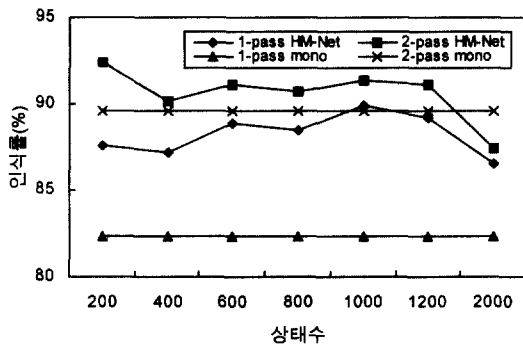



그림 7. 연속음성에 포함된 단어 인식률

때 HM-Net triphone의 경우 2-pass의 인식률은 평균 92.4%로서 단일 HMM에 비해 평균 2.8%의 향상된 인식률을 구하였다.

그리고 상태수의 증가에 따라 연속음성인식률과 단어인식률이 감소하는 원인으로는 학습에 참가한 음성 데이터의 부족으로 인해 정확한 HM-Net이 생성되지 못한 것으로 생각된다. 이는 향후 음향모델을 작성하는데 많은 양의 음성 데이터를 사용할 경우 해결할 수 있을 것으로 기대된다.

이상의 결과들로부터 본 연구에서 한국어에 적합한 문맥의존 음향모델을 작성하기 위해 적용한 한국어 음성학적 지식과 결정트리기반 상태분할 알고리즘의 유효성을 확인할 수 있었다.

다음은 인식시간을 줄이기 위한 프레임 단위 적응 프루닝 알고리즘을 연속음성인식에 대하여 적용한 결과를 표 3에 나타낸다. 앞의 실험결과에서 연속음성인식률 90.9%, 단어인식률 90.7%로 높은 인식률을 보인 상태수가 800에 대하여 적응 프루닝 알고리즘을 적용하였다. 본 연구에서는 일반적으로 널리 사용되는 고정 프루닝 문턱치와 비교를 하였다. 고정 프루닝 문턱치가 250인 경우에 인식률의 변화없이 가장 좋은 결과를 보였다.

표 3. 적응 프루닝 알고리즘을 적용한 실험 결과

고정 프루닝 문턱치		
연속음성인식률	단어인식률	인식시간
90.9%	90.7%	7.89초
적응 프루닝 문턱치		
연속음성인식률	단어인식률	인식시간
90.9%	90.7%	6.73초

위 표에서 보는 바와 같이 인식률의 변화없이 인식시간이 약 1.2초 줄어들어 유효성을 확인할 수 있었다. 또한, 기존의 고정 프루닝 알고리즘이 여러 번의 사전실험을 거쳐서 문턱치를 결정하는데 비해 적응 프루닝 알고리즘은 단어, 연결단어, 그리고 연속음성 등과 같이 다양한 태스크에도 사전실험이 필요없이 바로 적용할 수 있어 유효함을 확인할 수 있었다.

7. 결 론

본 연구에서는 연속음성인식 시스템의 성능향상을 위하여 인식률과 인식속도라는 두 가지 면에서 연구를 수행하였다. 먼저 인식률 향상을 위하여 효과적인 음향모델을 구성하기 위하여 PDT-SSS 알고리즘을 도입하여 HM-Net을 구성하였고, 언어모델로서 반복학습을 이용하여 인식실험을 수행한 결과 높은 인식률을 얻어 도입된 알고리즘의 유효성을 확인하였다. 그리고, 기존의 연결 단어 인식에서 유효성이 입증된 프레임 단위 적응 프루닝 문턱치 알고리즘을 연속음성에 적용하여 고정 프루닝 알고리즘에 비하여 인식시간이 줄어들어, 연속음성에서도 적응 프루닝 알고리즘의 유효성을 확인하였다.

PDT-SSS 알고리즘은 기존의 SSS 알고리즘에 한국어 음성학적 지식으로 구성한 음소질의어 집합을 이용한 음소 결정트리에 기반한 상태분할 방법으로 적절한 문맥 클래스의 분할과 임의의 문맥을 표현할 수 있다. 그리고, 프레임 단위 적응 프루닝 알고리즘은 이웃 프레임사이의 최대 확률의 상관성이 큰 점에 착안하여, 앞 프레임의 최대 확률로부터 효과적으로 프루닝 문턱치를 얻는 방법으로 현재 프레임에서 적응 프루닝 문턱치는 앞 프레임의 최대 확률과 후보 확률의 조합으로 결정할 수 있다.

제안된 방법의 유효성을 확인하기 위하여 항공편

예약 관련 연속음성인식 시스템에 적용하여 인식실험을 수행한 결과, 연속음성 인식률이 90.9%와 단어 인식률이 90.7%로 높은 인식률을 얻었으며, 적응 프루닝 알고리즘을 적용한 경우 인식률의 저하없이 고정 프루닝 알고리즘에 비해 인식시간이 약 1.2초(전체의 15%) 줄어들어 제안된 알고리즘의 유효성을 확인할 수 있었다.

참 고 문 헌

[1] 정현열, "음성인식 연구의 국내외 현황과 전망," 제15회 음성통신 및 신호처리 워크샵 논문집, pp. 23-30, 1998. 8.

[2] 김순협, "음성인식의 현황과 최근 연구 동향," 2000년도 한국음향학회 학술발표대회 논문집, Vol. 19, No. 2(s), 2000. 11.

[3] M. K. Ravishankar, "Efficient Algorithms for Speech Recognition," Ph. D Thesis, Carnegie Mellon University, 1996.

[4] Alleva, F., et al, "Applying SPHINX-II to the DARPA Wall Street Journal CSR task," Proc. of Speech and Natural Language Workshop, pp. 393-398, Feb. 1992.

[5] Alon Lavie, et al, "JANUS-III: Speech-to-speech translation in multiple languages," Proc. IEEE ICASSP-97, Vol.1, pp. 99-102, April 1997.

[6] A. Kai and S. Nakagawa, "A frame-synchronous continuous speech recognition algorithm using a top-down parsing of context-free grammar," Proc. ICSLP 92, pp. 257-260, 1992.

[7] T. Nishimoto, N. Shida, T. Kobayashi, K. Shirai, "Multimodal Drawing Tool Using Speech, Mouse and Keyboard," Porc. ICSLP, Vol. 3, pp. 1287-1290, 1994.

[8] Katsuhiko Shirai, "Spoken Dialogue in Multimodal Human Interface," ICSP '97, pp. 13-20, Aug. 1997.

[9] Hyun-Yeol Chung, Cheol-Jun Hwang, Shi-Wook Lee, "A Bimodal Korean Address Entry/Retrieval System," ICSLP98, pp. 1607~1610, 1998. 12.

[10] 김득수, 황철준, 정현열, "음성인식 기능을 가진 주소입력 시스템의 개발과 평가," 한국음향학회지 18권 2호, pp. 3-10, 1999. 2.

[11] 황철준, 오세진, 김범국, 정호열, 정현열, "실시간 주소인식을 위한 시스템의 인식속도 개선," 1999년도 한국음향학회 하계학술대회 논문집, 제18권 1(s)호, pp. 74-77, 1999. 7.

[12] Cheol-Jun Hwang, Se-Jin OH, Ho-Youl Jung, Hyun-Yeol Chung, "An Adaptive Pruning Threshold Algorithm for Efficient Speech Recognition," SPECOM '99, pp. 103-106, 1999. 10.

[13] 황철준, 오세진, 김범국, 정호열, 정현열, "음성인식의 고속화를 위한 프레임 단위 프루닝 알고리즘," 2000년도 한국음향학회 정기총회 및 학술발표대회 논문집, 제19권 2(s)호, pp. 183-186, 2000. 11.

[14] 오세진, 황철준, 김범국, 정호열, 정현열, "반복학습법에 의해 작성한 N-gram 언어모델을 이용한 연속음성인식에 관한 연구," 한국음향학회지 19권 6호, 2000. 8.

[15] 이호영, "국어음성학," 태학사, 1996.

[16] J. Takamia, S. Sagayama, "A Successive State Splitting Algorithm for Efficient Allophone Modeling," Proc. of ICASSP'92, pp. 573-576, 1992.

[17] 오세진, 임영춘, 황철준, 김범국, 정현열, "Hidden Markov Network를 이용한 음향학적 음소모델 작성에 관한 검토," 2000년도 한국음향학회 학술발표대회 논문집, 제19권 제2(s)호, pp. 29-32, 2000. 11.

[18] L.R. Bahl, P.V.de Souza, P.S. Gopalakrishnan, D. Nahamoo, M.A. Picheny, "Dicision Trees for Phonological Rules in Continuous Speech," Proc. of ICASSP'91, pp. 185-188, 1991.

[19] Se-Jin Oh, Chul-Joon Hwang, Bum-Kook Kim, Ho-Youl Jung, Hyun-Yeol Chung, "A Study on Speech Recognition using New State Clustering Algorithm of HM-Net with Korean Phonological Rules," Proc. of IC-AI '2001, U.S.A, 2001. 6.

[20] S. Furui, "Speaker-Independent Isolated Word

Recognition Using Dynamic Features of Speech Spectrum," IEEE Trans. Acoustics, Speech and Signal Processing, Vol. 34, No. 1, pp. 52-59, Feb. 1986.

[21] 坂井利, 中川聖一, "構文情報を用いた連續音聲認識," 情報處理學會 第15回 全國大會, pp. 37, Dec. 1994.

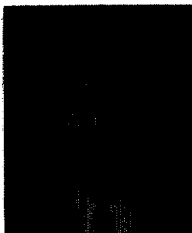
[22] B. T. Lowerre, "HARPY speech recognition system," Ph. D thesis, Carnegie Mellon University, 1976.

[23] J. C. Junqua, and J. P. Haton, "Robustness in Automatic Speech Recognition," Kluwer Academic Publishers, 1996.

[24] S. J. Young, P. C. Woodland, "State Clustering in hidden Markov model-based Continuous Speech Recognition," Computer Speech and Language, Vol. 8, No. 4, pp. 369-383, 1994.

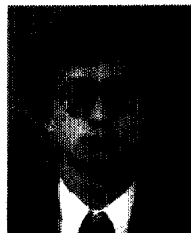
[25] 中川聖一, "確率モデ゛ルによる音聲認識," 日本電子情報通信學會, 1988.

황 철 준

1996년 2월 영남대학교 전자공학과(공학사)
1998년 2월 영남대학교 대학원 전자공학과(공학석사)
1998년 3월～현재 영남대학교 대학원 전자공학과(박사수료)
2000년 3월～현재 대구과학대학 정보전자통신계열 전임강사
관심분야：음성분석 및 인식, 디지털 신호처리

오 세 진

1996년 2월 영남대학교 전자공학과(공학사)
1998년 2월 영남대학교 대학원 전자공학과(공학석사)
1998년 3월～현재 영남대학교 대학원 전자공학과(박사수료)
2001년 9월～현재 대구과학대학 정보전자통신계열 전임강사
관심분야：음성분석 및 인식, 언어처리

김 범 국

1990년 2월 영남대학교 수학과(이학사)
1992년 2월 영남대학교 대학원 전자공학과(공학석사)
1998년 2월 영남대학교 대학원 전자공학과(공학박사)
1997년 3월～현재 대구과학대학 정보전자통신계열 조교수
관심분야：음성분석 및 인식, 언어처리, 멀티모달 시스템

정 호 열

1988년 2월 아주대학교 전자공학과(공학사)
1990년 2월 아주대학교 전자공학과(공학석사)
1993년 2월 아주대학교 전자공학과(박사수료)
1998년 (프)리옹국립응용과학원(INSA de Lyon) 전자공학전공(공학박사)
1998년 4월～1998년 12월 (프)CREATIS 박사후 과정
1999년 3월～현재 영남대학교 전자정보공학부 조교수
관심분야：음성·영상 신호처리, 인공지능, 디지털 워터마킹 등

정 현 열

1975년 2월 육군사관학교/영남대학교 전자공학과(학사)
1981년 8월 영남대학교 전자공학과(석사)
1989년 3월 동북대학교 정보공학과(공학박사)
1989년 3월～현재 영남대학교 전자정보공학부 교수
관심분야 : 디지털 신호처리, 문자인식, 음성인식